BOUNDING CAUSAL EFFECTS IN ECOLOGICAL INFERENCE PROBLEMS

By Alejandro Corvalan, Emerson Melo, Robert Sherman, and Matt Shum[1]

March, 2015.

Abstract

This paper is concerned with making causal inferences with ecological data. Aggregate outcome information is combined with individual demographic information from separate data sources to make causal inferences about individual behavior. In addressing such problems, even under the selection on observables assumption often made in the treatment effects literature, it is not possible to identify causal effects of interest. However, recent results from the partial identification literature provide the tightest upper and lower bounds on these causal effects. We apply these bounds to data from Chilean mayoral elections that straddle a 2012 change in Chilean electoral law from compulsory to voluntary voting. Aggregate voting outcomes are combined with individual demographic information from separate data sources to determine the causal effect of the change in the law on voter turnout. The bounds analysis reveals that voluntary voting decreased expected voter turnout, and that other causal effects are overstated if the bounds analysis is ignored.

Section 1: Introduction

Ecological inference (EI) problems are a class of data combination problems in which aggregate outcome information from one data source is combined with individual demographic information from a separate data source to make inferences about individual outcomes. The objectives of EI include description and prediction of individual behavior, as well as causal inference about individual behavior. King (1997) treats EI problems where the principal objective is description of individual behavior in political science applications. King, Rosen, and Tanner (2004) contain articles addressing all three objectives from a number of different fields, including political science, economics, and epidemiology.

Our paper is concerned with making causal inferences with ecological data. We apply new results from Fan et al. (2014a) in order to make causal inferences about individual behavior in an EI problem of substantive interest in political science.

In the absence of the data combination problem, standard results from the treatment effects literature can be used to perform straightforward counterfactual inference to determine causal effects. Specifically, if data on outcomes and covariates are observed in the same data set, then it is straightforward, under standard assumptions using known methods, to identify and consistently estimate the usual causal effects of interest, such as the average treatment effect (ATE) or the average treatment effect on the treated (ATT). For example, one could apply propensity score methods under a standard selection on observables assumption, as in Rosenbaum and Rubin (1983).

[1]Universidad Diego Portales, Cornell University, California Institute of Technology, and California Institute of Technology, respectively.

However, Fan et al. (2014a) show that these causal effects, even under the selection on observables assumption, cannot be identified when aggregate outcome data is combined with individual demographic data from separate sources. The information lost through aggregation precludes identification. However, these authors also establish upper and lower bounds on ATE and ATT which are valid under data combination. Moreover, these bounds are sharp, meaning that they are the tightest bounds possible under the maintained assumptions.

We apply these results to our ecological data to estimate bounds on causal effects of the change from compulsory to voluntary voting on turnout in Chilean mayoral elections. In this application aggregate turnout data must be combined with individual-level census data in order to make causal inferences about the effect of this policy change on voter turnout. We present bounds on ATE and show that voluntary voting decreased expected voter turnout. We also show that ATT is overstated by a standard difference analysis. For example, for Chile as a whole, the standard difference analysis estimates almost a 27% decrease in turnout for the voting-age population under the new law. The robust bounds analysis, on the other hand, estimates anywhere from a 15% decrease to 1.2% increase in turnout. We show that this pattern holds for many other subsets of the population: ignoring the bounds analysis results in an overstatement of the negative effect of the change from compulsory to voluntary voting on turnout for the voting-age population under the new law.

The rest of the paper is organized as follows. In Section 2, we introduce the ecological inference model considered in this paper, and discuss how to bound the causal effects of interest, using the results in Fan et al. (2014a). Section 3 describes the background of the change in Chilean voting law as well as the data we use for our analysis. Section 4 presents the results of our bounds analysis. Finally, in Section 5 we conclude and indicate directions for future work.

Section 2: The Ecological Inference Framework and Bounding Causal Effects

In this section, we introduce the ecological inference model considered in this paper. We define the causal effects of interest in this paper, namely, the average treatment effect ATE, and the average treatment effect on the treated ATT. Then, using the results in Fan et al. (2014a), we define sharp population bounds on ATE and ATT and show how to estimate these bounds. Finally, we discuss asymptotic results that allow inference on ATE and ATT.

Let D denote an observed binary treatment assignment indicator. That is, $D = 1$ if an individual is assigned to the treatment group and $D = 0$ if an individual is assigned to the control group. Let Y_D denote an individual outcome of interest. We adopt the "potential outcomes" approach to determining treatment effects pioneered by Rubin (1974). This approach views each individual as having a treatment outcome Y_1 and a control outcome Y_0, but only one of Y_1 and Y_0 is actually observed. Thus, the observed individual outcome is $Y = Y_1D + Y_0(1 - D)$. Let Z denote observed covariates which can effect both D and (Y_1, Y_0).

The standard "potential outcomes" approach requires that the analyst observe (Y, D, Z) for each individual in the sample. In the applications we consider, the link between observed outcomes and

covariates is considerably weaker. Instead of observing outcomes and covariates for each individual receiving a given treatment, we observe outcomes on one set of individuals who undergo a given treatment, and we observe covariates on a different set of individuals who receive that treatment. In other words, we observe *separate* outcome and covariate data sets. The outcome data set contains (Y, D) while the covariate data set contains (D, Z). Both data sets contain the treatment variable D which links the two sources of information. The objective is to combine these data sources to make inferences about the effect of treatment on outcomes. This is an ecological inference problem. We also note that in some applications, we observe separate outcome and covariate data sets for each treatment. That is, we observe $(Y_1, D = 1)$ and $(D = 1, Z)$, the treatment outcomes and covariates, in separate data sets. Likewise, we observe $(Y_0, D = 0)$ and $(D = 0, Z)$, the control outcomes and covariates, in separate data sets.

Next, we present the selection on observables and overlap assumptions, which are commonly made in the treatment effects literature (even without data combination) and which we also make. Selection on observables is a conditional independence assumption and overlap is a support assumption. Let $\mathcal{Z}$ denote the support of the covariate vector Z. For each $z \in \mathcal{Z}$ let $p(z) = I\!P\{D = 1 \mid Z = z\}$, the so-called propensity score.

A1. Selection on Observables: (Y_1, Y_0) is independent of D given $Z = z$.

A2. Overlap: For each $z \in \mathcal{Z}$, $0 < p(z) < 1$.

Randomized trials imply that (Y_1, Y_0, Z) is independent of D, which says that treatment and control outcomes, as well as observed covariates, are independent of treatment assignment. In fact, randomized trials imply that *no* variables are confounded with the treatment: the distribution of *all* variables, observed and unobserved, that affect treatment and control outcomes is the same in the treatment and control groups. In this sense, the only difference between treatment and control outcomes is the treatment, and so the causal effect of the treatment can be inferred from a comparison of treatment and control outcomes. The selection on observables assumption A1 is a *conditional* randomized trial assumption: once we condition on observables Z, outcomes are independent of treatment assignment. That is, given Z, there are no confounding variables: the distribution of all *unobserved* variables that affect outcomes is the same in the treatment and control groups. However, A1 allows *observed* variables to be confounded with the treatment in the sense that the distribution of observed variables is allowed to be different in the treatment and control groups. The overlap assumption A2 states that for each $z \in \mathcal{Z}$, there is a positive probability that some individual is assigned to the treatment group and a positive probability that some individual is assigned to the control group. Assumption A2 guarantees that in large samples there will be both treatment and control outcomes for each $z \in \mathcal{Z}$. Assumptions A1 and A2 make valid comparison of treatment and control outcomes possible for each $z \in \mathcal{Z}$.

When there is no data combination problem, that is, if (Y, D, Z) is observed for each individual in the sample, then under A1 and A2, standard propensity score methods (see, for example, Rosenbaum and Rubin (1983)) can be applied to point-identify and consistently estimate causal effects like ATE

and ATT. On the other hand, when there is a data combination problem – that is, when (Y, D, Z) are only observed in separate datasets – Fan et. al. (2014a) show that even if A1 and A2 hold, these causal effects cannot be identified. However, they go on to derive sharp upper and lower bounds on quantities like ATE and ATT using inequalities from the copula literature.

Recall the propensity score $p(Z) = I\!P\{D = 1 \mid Z\}$. Let $W = 1/p(Z)$ and $V = 1/[1 - p(Z)]$. Define $p_1 = I\!P\{D = 1\}$, the marginal probability of receiving treatment. Define $p_0 = 1 - p_1$. Foreshadowing our application, we develop notation for the special but common case in which the treatment and control outcomes Y_1 and Y_0, and therefore the observed outcomes Y, are binary.

Define $p_{00} = I\!P\{Y = 0 \mid D = 0\}$, $p_{01} = I\!P\{Y = 0 \mid D = 1\}$, and $p_{11} = I\!P\{Y = 1, D = 1\}$. Let X denote an arbitrary random variable. For $d = 0, 1$, write $F_{X|D}(\cdot \mid d)$ for the cumulative distribution function of X given $D = d$. Write $Q_{X|D}(\cdot \mid d)$ for the quantile function of X given $D = d$. Define the average treatment effect ATE and the average treatment effect on the treated ATT as follows:

$$\begin{aligned} ATE &\equiv I\!E(Y_1 - Y_0) = I\!P\{Y_1 = 1\} - I\!P\{Y_0 = 1\} \\ ATT &\equiv I\!E(Y_1 - Y_0 \mid D = 1) = I\!P\{Y_1 = 1 \mid D = 1\} - I\!P\{Y_0 = 1 \mid D = 1\}. \end{aligned}$$

The following result is a special case of Theorem 3.2 in Fan et al. (2014a).

THEOREM 1. *Suppose $Var(X) < \infty$ and $Var(V) < \infty$. If A1 and A2 hold, then*

$$\begin{aligned} \mu_1^L - \mu_0^U \leq \quad & ATE \quad \leq \mu_1^U - \mu_0^L \\ p_{11}/p_1 - \mu_{0|1}^U \leq \quad & ATT \quad \leq p_{11}/p_1 - \mu_{0|1}^L \end{aligned}$$

where

$$\begin{aligned} \mu_1^L &= p_1 \int_0^{p_{01}} Q_{W|D}(u \mid 1) du \\ \mu_1^U &= p_1 \int_{p_{01}}^1 Q_{W|D}(u \mid 1) du \\ \mu_0^L &= p_0 \int_0^{p_{00}} Q_{V|D}(u \mid 0) du \\ \mu_0^U &= p_0 \int_{p_{00}}^1 Q_{V|D}(u \mid 0) du \\ \mu_{0|1}^L &= \frac{p_0}{p_1} \int_0^{p_{00}} Q_{V/W|D}(u \mid 0) du \\ \mu_{0|1}^U &= \frac{p_0}{p_1} \int_{p_{00}}^1 Q_{V/W|D}(u \mid 0) du. \end{aligned}$$

Let (Y_i, D_i), $i = 1, \ldots n_1$ denote iid observations of outcome and treatment variables from the outcome data set(s). Let (D_j, Z_j), $j = 1, \ldots, n_2$ denote iid observations of treatment and demographic variables from the covariate data set(s). We estimate the population intervals with corresponding sample intervals:

$$\left[\hat{\mu}_1^L - \hat{\mu}_0^U \,,\, \hat{\mu}_1^U - \hat{\mu}_0^L\right] \tag{1}$$

$$\left[\hat{p}_{11}/\hat{p}_1 - \hat{\mu}_{0|1}^U \,,\, \hat{p}_{11}/\hat{p}_1 - \hat{\mu}_{0|1}^L\right] \tag{2}$$

where

$$\begin{aligned}
\hat{\mu}_1^L &= \hat{p}_1 \int_0^{\hat{p}_{01}} \hat{Q}_{W|D}(u \mid 1)du \\
\hat{\mu}_1^U &= \hat{p}_1 \int_{\hat{p}_{01}}^1 \hat{Q}_{W|D}(u \mid 1)du \\
\hat{\mu}_0^L &= \hat{p}_0 \int_0^{\hat{p}_{00}} \hat{Q}_{V|D}(u \mid 0)du \\
\hat{\mu}_0^U &= \hat{p}_0 \int_{\hat{p}_{00}}^1 \hat{Q}_{V|D}(u \mid 0)du \\
\hat{\mu}_{0|1}^L &= \frac{\hat{p}_0}{\hat{p}_1} \int_0^{\hat{p}_{00}} \hat{Q}_{V/W|D}(u \mid 0)du \\
\hat{\mu}_{0|1}^U &= \frac{\hat{p}_0}{\hat{p}_1} \int_{\hat{p}_{00}}^1 \hat{Q}_{V/W|D}(u \mid 0)du \,.
\end{aligned}$$

We use (Y_i, D_i), $i = 1, \ldots, n_1$ to construct the sample proportions $\hat{p}_1$, $\hat{p}_0$, $\hat{p}_{01}$, $\hat{p}_{00}$, and $\hat{p}_{11}$. For example, $\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\{D_i = 1\}$, $\hat{p}_{01} = \frac{1}{n_1\hat{p}_1}\sum_{i=1}^{n_1}\{Y_i = 0, D_i = 1\}$, $\hat{p}_{11} = \frac{1}{n_1}\sum_{i=1}^{n_1}\{Y_i = 1, D_i = 1\}$, and so on.

We use (D_j, Z_j), $j = 1, \ldots, n_2$ to construct $\hat{p}(Z)$, a consistent estimator of the propensity score. There are many ways to estimate the propensity score. One can use parametric estimation procedures like probit or logit, semiparametric estimation procedures, or nonparametric estimation procedures. The estimated quantile functions above are functions of the estimated quantile function of the propensity score. For ease of notation, define $P = p(Z)$. For $d = 0, 1$, we define the estimated quantile function of P given $D = d$ to be $\hat{Q}_{P|D}(u \mid d) = \inf\{a : \hat{F}_{P|D}(a \mid d) > u\}$ where $\hat{F}_{P|D}(\cdot \mid d)$ is the estimated empirical cumulative distribution function of P given $D = d$. That is, with $\hat{P}_j = \hat{p}(Z_j)$, $\hat{F}_{P|D}(a \mid d) = \frac{1}{n_2\hat{p}_d}\sum_{j=1}^{n_2}\{\hat{P}_j \leq a, D_j = d\}$. Using the fact that W is a monotone decreasing function of P, and V and V/W are monotone increasing functions of P, we get that

$$\begin{aligned}
\hat{Q}_{W|D}(u \mid d) &= 1/\hat{Q}_{P|D}(1 - u \mid d) \\
\hat{Q}_{V|D}(u \mid d) &= 1/[1 - \hat{Q}_{P|D}(u \mid d)] \\
\hat{Q}_{V/W|D}(u \mid d) &= \hat{Q}_{P|D}(u \mid d)/[1 - \hat{Q}_{P|D}(u \mid d)] \,.
\end{aligned}$$

Finally, the integrals in the expressions above are numerical integrals over the indicated subsets of the unit interval.

Theorems 6.1 and 6.2 in Fan et al. (2014b) can be used to prove that the vector of lower and upper bound estimators for both ATE and ATT are $\sqrt{n}$-consistent and jointly asymptotically normally distributed estimators of their population counterparts. This result permits us to apply the methods in Stoye (2009) to compute asymptotic confidence intervals for ATE and ATT.

From the time democracy was reintroduced in Chile in 1989 until the registration and voting system was reformed in 2012, electoral participation required self-initiated registration into the electoral rolls. Notably, registration was voluntary while voting, conditional on being registered, was compulsory. This combination is rare in the world. Most countries have either automatic registration with voluntary voting (e.g., Germany, Britain), both voluntary registration and voluntary voting (the United States), or automatic registration and compulsory voting (Belgium).

In January of 2012 the Chilean government passed a law replacing a system of compulsory voting in municipal, parliamentary, and presidential elections, with a system of voluntary voting. Prior to the 2012 reforms, the cost of registration in Chile was particularly high. Citizens had to register in person at the registration office in the district where they wanted to vote. There were few registration offices in electoral districts, and most were not located near easily accessible bureaucratic offices. There were also significant information and scheduling costs.[2] In addition, the mixture of voluntary registration and compulsory voting further increased the cost of registration for new entrants. According to the 2010 Chilean National Youth Survey, one out of ten nonregistered citizens did not register because of the perceived burden of having to vote in all subsequent elections.[3] Part of this burden was the threat of a substantial monetary penalty if one failed to vote.

The effects of the Chilean voluntary registration and compulsory voting policy on the age structure of voters are notorious. Older citizens, who registered in large numbers in 1988 in order to vote in the first democratic referendum, were generally committed voters, conforming to the mandatory feature of the rule. Younger cohorts, however, had been increasingly reluctant to register during the post-authoritarian period. As a result, older voters were over-represented while younger voters were under-represented: the turnout rate for those aged 35 and above was close to 90 percent in 2009, while in the same year the rate for adults aged $18-29$ was only 23 percent. Carlin (2006) and Corvalan and Cox (2013) note that this participation rate for younger voters was by far the lowest among Latin American countries. Indeed, the desire to increase voter turnout among young voters was a principal motive to reform the registration and voting rules. In January 31, 2012, Electoral Law 20,568 made registration automatic and voting voluntary.

Since the passage of these reforms in 2012, all eligible voters are automatically registered to vote and may voluntarily vote in presidential, parliamentary, and municipal elections. A natural question to ask is "What effect did the change in the law have on election turnout?" We address this causal question using the new methods. Since the first elections under the new system were the municipal elections in 2012, we focus our analysis on the most important of the municipal races, namely, the races for mayor.

[2]These costs are consistent with findings in Rosenstone and Wolfinger (1978). It should also be noted that a registered voter could cancel his or her registration. However, the cancellation process was just as burdensome as registration itself.

[3]These survey results are consistent with results on cost-benefit analysis in the theory of rational voting, as in Downs (1957) and Riker and Ordeshook (1968).

Chile is divided into 15 regions, which are subdivided into communes or counties. Each commune is governed by a municipality headed by a mayor and a municipal council. Municipal elections in Chile have taken place every four years since 1992. In each election, both the mayor and the council members are elected. Since 2004, the mayor has been elected separately from the council members. Mayoral candidates compete for one seat in each commune and are elected under plurality rule.

We have aggregate voting data for the first mayoral elections under voluntary voting in 2012 as well as aggregate voting data for the first direct mayoral election in 2004, when voting was still compulsory. Our source of voting data is INE, the Chilean National Statistics Office. Our source of covariate data is CASEN, the most complete Chilean socioeconomic survey. This survey is conducted by the Chilean government every two to three years in all the communes in the country. Unlike the aggregate voting data from INE, the CASEN data is individual-level data. The CASEN data is not aggregated at any level. Corresponding to aggregate voting data in the 2004 election, we use the 2003 CASEN survey, with a sample size of 257,077. Corresponding to aggregate voting data in the 2012 election, we use the 2011 CASEN survey, with a sample size of 200,302. We consider data for those individuals at least 18 years old, the minimum voting age.

Section 4: Data Analysis and Results

A naïve measure. A naïve measure of the causal effect of the new voting law on turnout is the simple difference between turnout proportions in 2012 and 2004. This measure is an unbiased estimate of the causal effect of the change from compulsory to voluntary voting only if there are no confounding variables, which means that the distribution of all observed and unobserved variables affecting turnout is the same in both election years. But this is implausible. For example, the distribution of household income is different in 2004 and 2012, and income is likely to affect turnout. That is, income is likely to be a confounding factor. Age may also be a confounding factor. As mentioned in Section 3, the change to voluntary voting was motivated in part by the desire to increase turnout among young voters. Moreover, even if a strong exogeneity condition holds, standard linear and binary regression methods are impracticable because of the data combination problem. Similar objections can be raised about simple difference-in-differences methods as well as standard linear and binary regression versions of the difference-in-differences techniques.

A new approach. Given the inadequacies of the naïve difference measure, we turn to our new approach. In the Chilean voting application, the treatment $D = 1$ corresponds to voluntary voting in Chilean mayoral elections in 2012 and the treatment $D = 0$ corresponds to compulsory voting in Chilean mayoral elections in 2004. The treatment outcome Y_1 is a binary outcome equal to unity if, under voluntary voting in 2012, an eligible voter turns out to vote, and zero otherwise. The control outcome Y_0 is a binary outcome equal to unity if, under compulsory voting in 2004, an eligible voter

turns out to vote, and zero otherwise.[4] The observed outcome is $Y = Y_1 D + Y_0(1-D)$.[5] As explained in detail in Section 3, the treatment outcomes $(Y_1, D = 1)$ are obtained from a 2012 data source while the treatment covariates $(D = 1, Z)$ are obtained from a separate 2011 data source. Similarly, the control outcomes $(Y_0, D = 0)$ are obtained from a 2004 data source while the control covariates $(D = 0, Z)$ are obtained in a separate 2003 data source. Accordingly, $ATE = I\!P\{Y_1 = 1\} - I\!P\{Y_0 = 1\}$ and $ATT = I\!P\{Y_1 = 1 \mid D = 1\} - I\!P\{Y_0 = 1 \mid D = 1\}$.

The assumptions underlying our approach have natural interpretations in the Chilean voting application. Specifically, assumption A1 says that conditional on observed covariates like income and age, no unobserved variables are confounded with the change from compulsory to voluntary voting. In other words, given observed covariates, the distribution of all unobserved variables that affect turnout decisions in mayoral elections is the same in 2004 under compulsory voting as it is in 2012 under voluntary voting. For example, one of the unobserved variables in our model that may affect turnout decisions is mayoral candidate quality. Assumption A1 states that conditional on observed covariates, the distribution of mayoral candidate quality (as well as other unobserved variables that affect turnout decisions) is the same in 2004 as it is in 2012. However, assumption A1 allows the effect of observed covariates on turnout to be confounded with the effect on turnout of the change from compulsory to voluntary voting. We interpret the propensity score $p(z)$ as the probability, conditional on $Z = z$, that an observation comes from 2012 rather than 2004. Assumption A2 says that for each possible value of the vector of observed covariates, there is a positive probability that an eligible voter in 2012 makes a turnout decision and there is a positive probability that an eligible voter in 2004 makes a turnout decision.

Under Assumptions A1 and A2, ATE is the average change in turnout in mayoral elections in 2012 relative to 2004 due to the change from compulsory to voluntary voting, while ATT is the average change in turnout in these elections due to the change in voting laws for those eligible to vote in 2012. Since the current law makes registration automatic, ATT is arguably just as interesting a causal measure as ATE.

In this section, we present estimated bounds on ATE and ATT for the entire population of Chile as well as for interesting subsets of this population, such as the population of men, the population of women, the 15 regions of Chile, and different age groups.

As stated previously, in this application, the observed outcome $Y = Y_1 D + Y_0(1 - D)$ where Y_1 is an indicator of a turnout decision made in the mayoral election in 2012 by an eligible voter after the change from compulsory to voluntary voting, Y_0 is an indicator of a turnout decision made in the mayoral election in 2004 by an eligible voter before the change from compulsory to voluntary voting, and D is the indicator of the election year, where $D = 1$ if the election year is 2012 and $D = 0$ if the

[4] For convenience, we use registration as a proxy for voting in 2004. Since voting is compulsory in 2004, the differences between those who register and those who vote in 2004 is very small. Also, note that under compulsory voting, $Y_0 = 0$ if an eligible voter is not registered, since registration is a necessary condition for voting.

[5] Note that an eligible voter in 2004 can also be an eligible voter in 2012. However, D can never be both zero and one since our methods formally treat each voter in 2004 as different from each voter in 2012.

election year is 2004. Note that we are identifying $D = 1$ with the treatment voluntary voting and $D = 0$ with the control compulsory voting. This identification is valid under assumption A1.

The treatment outcomes $(Y_1, D = 1)$ are obtained from a 2012 INE data set while the treatment covariates $(D = 1, Z)$ are obtained from a 2011 CASEN data set. Similarly, the control outcomes $(Y_0, D = 0)$ are obtained from a 2004 INE data set while the control covariates $(D = 0, Z)$ are obtained from a 2003 CASEN data set. We take the observed covariate zector $Z = (Z_1, \ldots, Z_6) = (loginc, age, educ, gender, unemp, married)$. Table 1 describes each component of Z and gives corresponding summary statistics for Chile in both 2003 and 2011.

For a given population subset of interest, let (Y_i, D_i), $i = 1, \ldots n_1$ denote observations of outcome and treatment variables from the combined INE outcome data sets from 2004 and 2012, and let (D_j, Z_j), $j = 1, \ldots, n_2$ denote observations of treatment and demographic variables from the combined CASEN covariate data sets from 2003 and 2011. For the given subset of interest, n_1 is sample size of the combined INE outcome data sets and n_2 is the sample size of the combined CASEN covariate data sets.

In order to estimate the bounds on ATE and ATT given in Theorem 1, we must first estimate the propensity score $p(Z_j) = I\!P\{D_j = 1 \mid Z_j\}$ using the CASEN data (D_j, Z_j), $j = 1, \ldots, n_2$ from each population subset of interest. For the country as a whole and for each of the 15 regions of Chile we estimate the propensity score by estimating the coefficients of the probit regression

$$I\!P\{D_j = 1 \mid Z_j\} = \Phi(\beta_0 + \beta_1 Z_{1j} + \beta_2 Z_{2j} + \beta_3 Z_{3j} + \beta_4 Z_{4j} + \beta_5 Z_{5j} + \beta_6 Z_{6j}) .$$

We also estimate separate propensity score models for men and women and separate models for age categories $18-24$, $25-29$, $30-34$, $35-39$, $40-44$, $45-49$, $50-54$, $55-59$, $60-64$, $65-69$, and $70-74$. The separate models for men and women have the same form as the probit regression above, except that the gender variable Z_4 is dropped from the model. Similarly, the probit regression for the separate age categories has the same form except that the age variable Z_2 is dropped.

We present results of estimating the probit regression above for the entire country. Recall that we use the combined CASEN data sets from 2011 and 2003 to estimate the propensity score. We interpret the propensity score as the conditional probability that an observation comes form 2012 rather than 2004.

Table 2 presents coefficient estimates and standard deviations for the probit regression for the entire country. We see that all the variables make a statistically significant marginal contribution to the model. The first five make positive contributions, whereas the married variable makes a negative contribution. We can interpret the significant positive coefficient on *loginc* as implying that ceteris paribus, eligible voters in 2012 had higher income than in 2004 (which reflects the overall rise in the economic fortunes of Chile as a whole during that period). Similar interpretations can be made for the other variables in the model.

It is useful at this point to make the following observation. Suppose that the set of observed covariates Z has no predictive power in the probit regression above. Note that this holds if D is independent of Z. Now, if D is independent of Z and A1 holds, then (Y_1, Y_0, Z) is independent of D,

which follows from a randomized trial assumption. Under the randomized trial assumption, ATE and ATT are equal and the simple difference estimator $\frac{\sum_{i=1}^{n_1} Y_i D_i}{\sum_{i=1}^{n_1} D_i} - \frac{\sum_{i=1}^{n_1} Y_i(1-D_i)}{\sum_{i=1}^{n_1}(1-D_i)}$ is a consistent estimator of ATE and ATT. No bounds analysis is needed in this case.

Now turn to Figure 1. Displayed in this figure are 95% confidence intervals for ATE and ATT for the country as a whole, as well as for men and women separately. For better visual effect, these confidence intervals are represented as boxes, where the length of a box is the length of the corresponding confidence interval. Focus on the box for ATE for the country. The ordinate of any point on the top of this box is the upper bound estimate for ATE given in (1) plus a standard error correction computed using the procedure of Stoye (2009). The ordinate of any point on the bottom of this box is the lower bound estimate for ATE given in (1) minus a standard error correction computed using Stoye's procedure.[6] The box is split in the middle by a line. The starred point represents the simple difference estimate defined in the last paragraph.[7] Corresponding statements apply to the other ATE boxes and to the ATT boxes.

Consider the ATE boxes in Figure 1. We see that for the country as a whole as well as for men and women separately, the simple difference estimates suggest that voluntary voting decreased voter turnout. This suggestion is confirmed by the robust bounds analysis: each 95% confidence interval upper bound is below the zero level.

Next consider the ATT boxes in Figure 1 and recall that ATT may be the more relevant causal measure since registration and therefore eligibility is automatic under current Chilean law. We see that the robust bounds do not contain the simple difference estimates. Under A1, this is strong evidence against the randomized trial assumption and strong evidence for the need for this type of bounds analysis. If the bounds analysis were ignored, the negative effect of voluntary voting on turnout for eligible voters in 2012 would be overstated. In fact, notice that all three ATT boxes contain the point zero, although just barely. This suggests that under assumptions A1 and A2, we cannot reject the hypothesis that voluntary voting had no effect on turnout for eligible voters in 2012 at the 5% level. On the other hand, this hypothesis might be rejected at a less stringent significance level.

Now consider Figure 2, which displays bounds results for ATE and ATT for the 15 regions comprising Chile. As a reference point, the last box in Figure 2 represents the results in Figure 1 for the country as a whole. The results for the individual regions are qualitatively the same as those for the country as a whole. All the ATE boxes are below the zero level and, with the exception of Region 15, contain the corresponding simple differences estimates. Note that the ATT boxes for Regions 5, 11, 12, and 13 are all below the zero level, implying that voluntary voting has, at the 5% level, a statistically significant negative effect on turnout in these regions. In all regions except possibly Regions 1 and 13,

[6] As mentioned in Section 2, the procedure of Stoye (2009) is valid under joint asymptotic normality of the lower and upper bound estimators given in (1) and (2). The joint asymptotic normality results are given in Theorem 6.1 and Theorem 6.2, respectively, in Fan et al. (2014b). The asymptotic standard errors in these theorems are estimated with the bootstrap to produce our standard error corrections. We note that the standard error corrections in this application are typically negligible compared to the length of the bounds.

[7] The turnout difference estimates can be taken as exact population differences. The reason is that they are based on very large sample sizes, making the length of the corresponding confidence intervals zero for all practical purposes.

ignoring the bounds analysis and taking the simple difference estimates at face value overstates the negative effect of voluntary voting on turnout for those eligible to vote in 2012.

Finally, consider Figure 3, which displays results for ATE and ATT conditional on age. The results are qualitatively similar to those presented in the previous figures. However, focus on the two youngest age categories, and recall that one of the motivations for changing from compulsory to voluntary voting was to try to increase turnout among young voters. We see that the ATE and the ATT boxes both straddle the zero level for the 18 - 24 and 25 - 29 age categories, and the ATT box for the 18 - 24 age category is nearly above the zero level. While not conclusive at the 5% significance level, the results do not rule out the possibility that voluntary voting had a positive effect on turnout among younger voters, in line with the intended goals of the policy change.

Section 5: Conclusion

This paper uses new partial identification results from the treatment effects literature on data combination to make inferences about causal effects in ecological inference problems. Of course, the need for causal inference and counterfactual evaluation (in contrast to simple before-after comparison of outcomes from policy changes) is well understood in political science, and methods are readily available. But these methods break down when the researcher must combine aggregate and individual-level data sources as part of the causal inference exercise. The novel contribution of this paper is to propose methodology which works in this case. More broadly, the need to combine different data sources in causal effect modelling appears commonplace in political science. Besides the application considered in this paper, other potential applications include measuring the effect of introducing electronic voting on vote outcomes, the effects of war on health outcomes, or the effects of political turmoil on economic activity. In all these cases, one needs to combine aggregate (precinct-, regional-, or country-level) outcome data with demographic confounders measured at the individual level.

We apply our methodology to bound causal effects of a change from compulsory to voluntary voting on turnout in recent Chilean mayoral elections. The bounds analysis reveals that the change had a negative effect on expected turnout and that ignoring this analysis and applying a simple difference estimator leads to overstating the negative effect of the change on those who are eligible to vote under the current voluntary voting laws. In future work, we plan to study the effect of the change in the law on turnout as well as other voting outcomes in recent parliamentary and presidential elections in Chile.

REFERENCES

Carlin, R. (2006): "The decline of citizen participation in electoral politics in post-authoritarian Chile," *Democratization*, **13**, 632–651.

Corvalan, A. and Cox, P. (2013): "Classbiased electoral participation: the youth vote in Chile," *Latin American Politics and Society*, **55**, 47–68.

Downs, A. (1957): "An economic theory of political action in a democracy," *The Journal of Political Economy*, 135–150.

Fan, Y., Sherman, R., and Shum, M. (2014a): "Identifying treatment effects under data combination," *Econometrica*, **82**, 811–822.

Fan, Y., Sherman, R., and Shum, M. (2014b): "Estimation and inference in an ecologocal inference model," Working paper.

King, G. (1997): *A Solution to the Ecological Inference Problem*, Princeton, Princeton University Press.

King, G., Rosen, O., and Tanner, M. (2004): *Ecological Inference: New Methodological Strategies*, Cambridge, Cambridge University Press.

Riker, W. and Ordeshook, P. (1968): "A Theory of the Calculus of Voting" *American Political Science Review*, **62**, 25–42.

Rosenbaum, P. and Rubin, D. (1983): "The central role of the propensity score in observational studies of causal effects," *Biometrica*, **70**, 41–55.

Rosenstone, S. and Wolfinger, R. (1978): "The effect of registration laws on voter turnout," *The American Political Science Review*, 22–45.

Rubin, D. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, **66**, 688–701.

Stoye, J (2009): "More on confidence intervals for partially identified parameters," *Econometrica*, **77**, 1299–1315.

Table 1: summary statistics for the country

		2003		2011	
variable	description	mean	std.dev	mean	std.dev
loginc	log of annual household income	11.10	1.17	11.86	1.16
age	in years	42.72	17.27	44.31	17.99
educ	completed years of schooling	8.65	4.43	10.07	4.33
gender	1 if female	.51	.49	.53	.49
unemp	1 if unemployed	.05	.22	.04	.20
married	1 if married	.60	.49	.55	.49
sample size		173,625		144,428	

Table 2: estimated propensity score model coefficients for the country

variables	coeff	std.dev
loginc	.32	.002
age	.008	.0002
educ	.034	.0006
gender	.074	.005
unemp	.12	.01
married	-.15	.005
n_2	318,053	

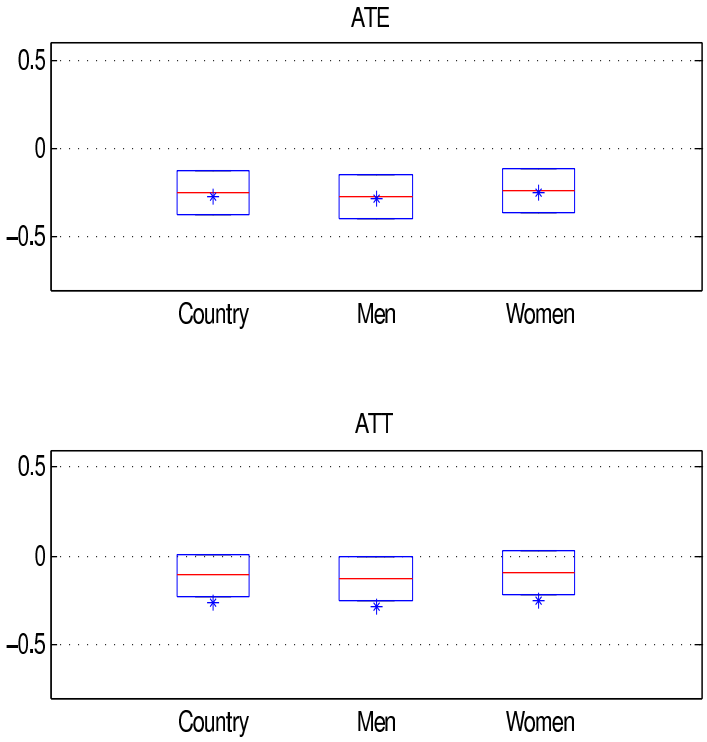
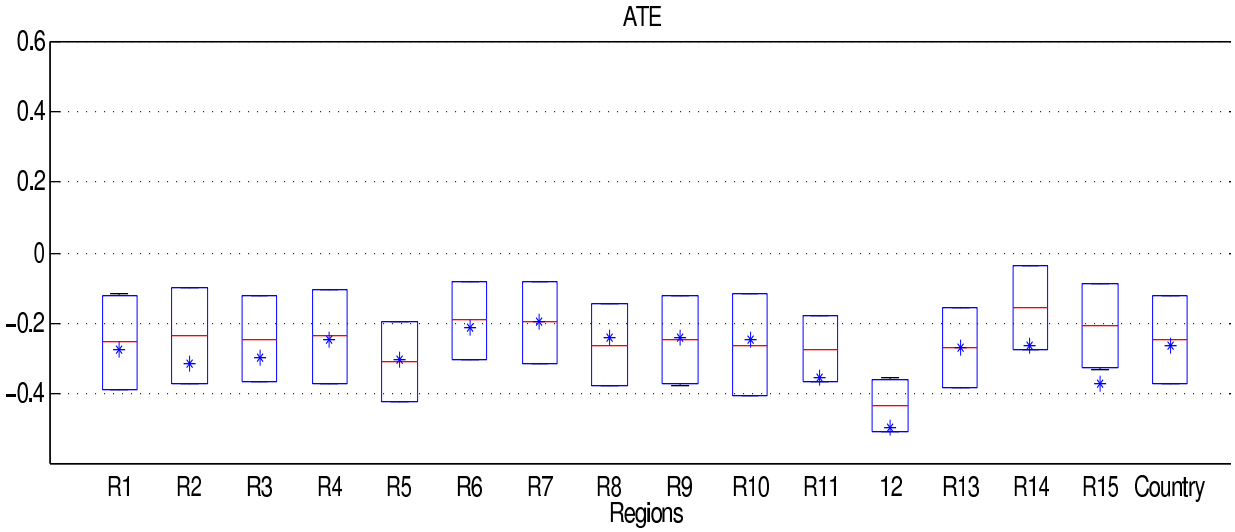


Figure 1:


ATE
0.6
0.4
0.2
0
−0.2
−0.4
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
12
R13
R14
R15
Country
Regions

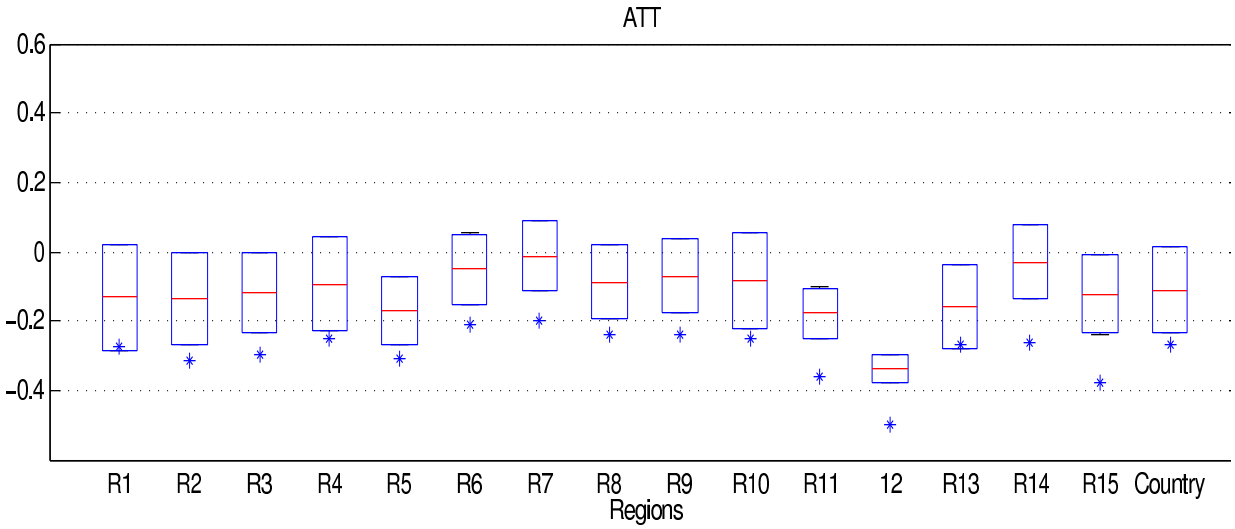
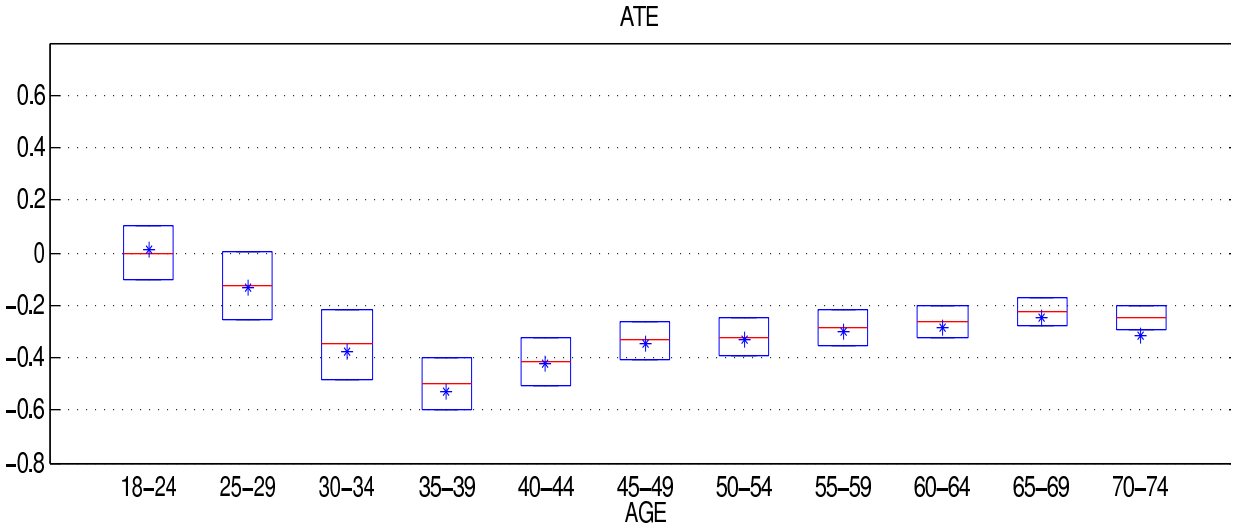

ATT
0.6
0.4
0.2
0
−0.2
−0.4
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
12
R13
R14
R15
Country
Regions

Figure 2:


ATE
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
18–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
AGE

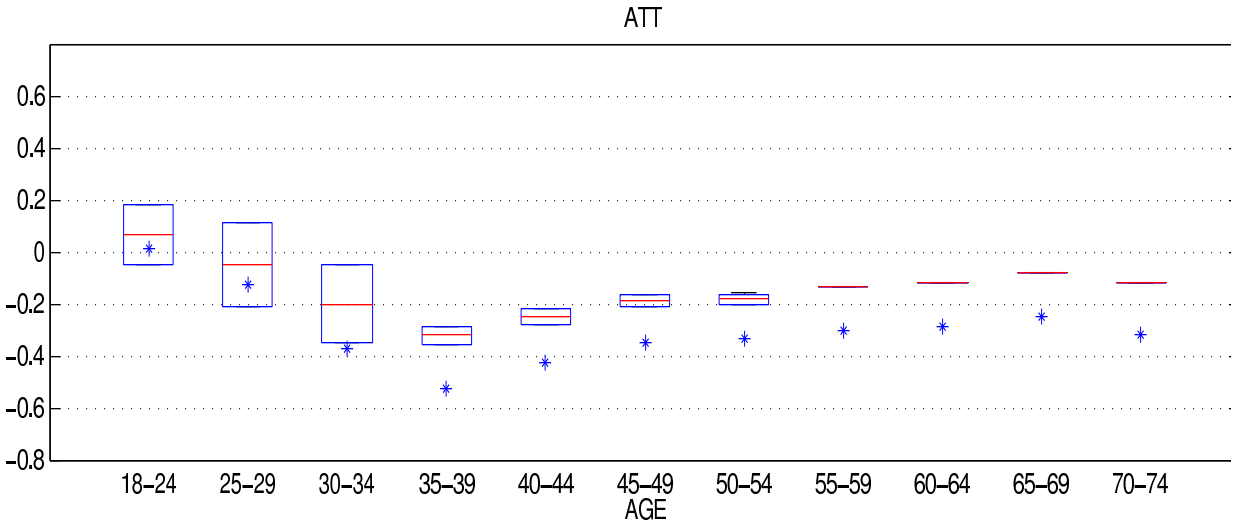

ATT
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
18–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
AGE

Figure 3: